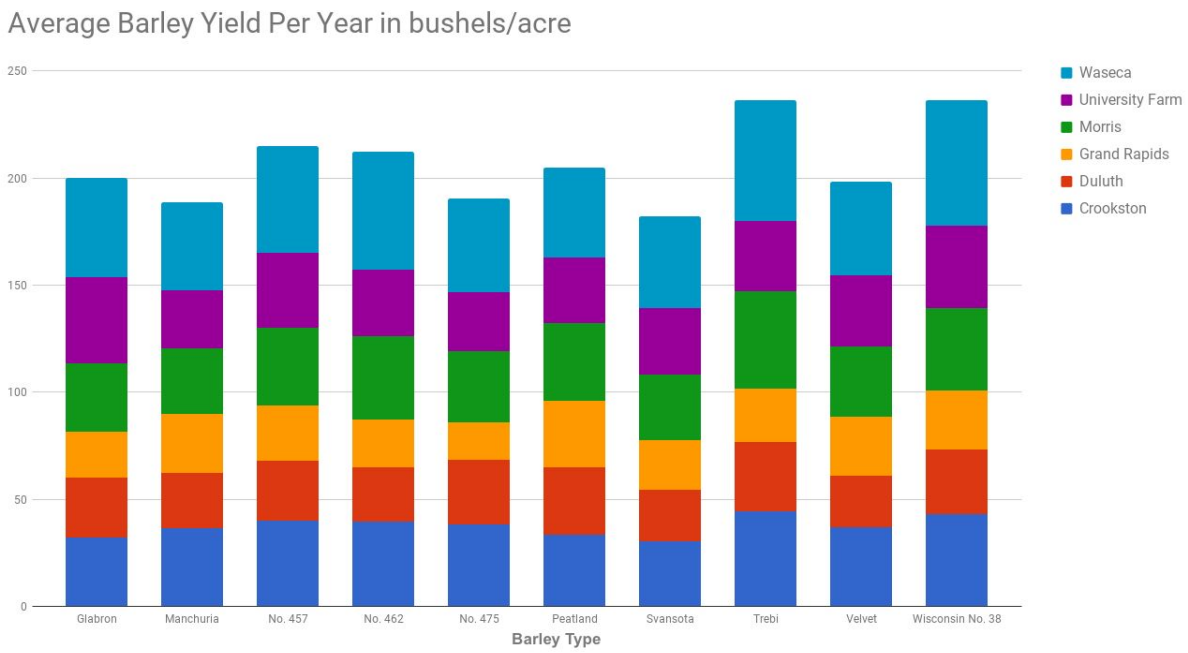
Cristian Lara (CLARAL3)
CS 448B - Assignment 1



There are 4 variable names in the data set, but I believe that there are two essential variables which the visualization should clearly reveal information about -- which varieties of barley have the highest yield. In order to emphasize these two variables the most, I dedicated two axes to them. The x-axis represents the different kinds of barley, while the y-axis represents the yield

If I only wanted to represent those two variables, then I could just sum up all the yields in the data for each type of barley (ignoring site and year) and use a scatter or bar graph, but there is still more interesting information in the dataset, namely, the sites where the barley was grown (what if a particular site had a very high yield?). I partitioned the data into 6 different tables, one for each different site, and averaged the rows representing yields for different years. This allowed me to create a stacked bar graph with nested bars indicating each site's contributions to a barley type's total yield across all the sites, while still being able to keep the total yield per berley type the focus of the visual.

I chose this type of graph because each barley type's yield is a sum of yields from many different tests, and there is valuable information in the breakdown of how each site contributed to the total yield. So there is an element of percentage that is important, which seems to encourage the use of a pie chart, but each sum must also be compared to each other, making the stacked bar graph a better option.

To better emphasize the data that I thought was important, I decided to not include any distinction between the two years that the data was gathered. I considered adding a third axis to represent the two years of tests, but it de-emphasized the main information and the different years really aren't the focus of the tests, they were more of a means to get more accurate results.